Research Proposal

On

Developing a novel algorithm using Machine learning for Cancer prediction

Submitted by

Poonam A. Chaudhari

Symbiosis Institute of Technology, Pune

Under the Guidance of

Dr. Himanshu Agarwal
Associate Professor
Symbiosis Institute of Technology, Pune

Faculty: Engineering

Symbiosis International University, Pune

Research Proposal Approval

The research proposal titled “Developing a novel algorithm using Machine learning for Cancer prediction” submitted by **Symbiosis Institute of Technology, Pune** (2000722) is prepared under my supervision and approved for further research work.

Guide

Dr. Himanshu Agarwal

Associate Professor

Symbiosis Institute of Technology, Pune

I declare that this research proposal has been prepared by me.

Poonam A. Chaudhari

SIRI ID: 2000722

1. INTRODUCTION

Cancer is one of the deadliest diseases affecting today. Cancer is a group of diseases characterized by uncontrolled cell division leading to growth of abnormal tissue causing tumors. [1] However, some tumours are benign and some are malignant. These malignant tumours affect the surrounding tissues and organs.

Cancer is ghastly as it often doesn't announce itself until it is at an advanced stage. There is no single test that can accurately diagnose cancer. Diagnosis of such a ruthless disease cannot be done only through clinical trials. It requires capability to deal with complex proteomic and genomic measurements. Thus, machine learning is frequently used in cancer diagnosis and detection.

Machine learning [2] is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. Machine Learning is widely classified as Supervised Learning, Unsupervised Learning, Reinforcement Learning and Deep Learning. Lot of research has been done in supervised and unsupervised learning algorithms like Linear Regression, Decision Trees, Random Forest, Support Vector Machine, etc.

We propose to define an algorithm which would make use of ensemble algorithms in deep learning to select a bio marker for lung cancer. This bio marker will find a specific gene expression from the miRNA sequence. We will generate trained data by mapping normal cells with tumour cells. This will help us classify test data as benign or malignant. Thus, predict the occurrence of the disease.

2. LITERATURE REVIEW

Initially, most applications of machine learning in the medical domain of cancer treatment were concerned with the diagnosis and detection of cancer, rather than with the response to treatment and prediction of outcome [3]. This was the driving force for Stiphout and group to explore a new area of prediction of cancer using machine learning. The age, gender, surgery type, residual group were few parameters considered for the same. Standard statistical method ie.Linear regression and machine learning method ie. Support Vector Machine were compared. The results showed no considerable difference. This created a scope of using feature selection for improving the performance, thus, emphasizing classification.

[4] Classification of correct and incorrect instances led to enhanced performance. Naive Bayes algorithm is based on probability and J48 algorithm is based on decision tree. Prof. Tina worked towards maximizing true positive rate and minimizing false positive rate. She created a confusion matrix with all combinations between true positive, false positive, precision and recall. Decision tree yielded a better performance as compared to probabilistic model.

[5] John Halloran worked towards showing the importance of size of training set for algorithm selection. Naive Bayes reaches its asmyptotic error very quickly with regards to the number of training examples. Thus, if training data is scarce, one can expect naive Bayes to outperform logistic regression, but as the number of training examples grows, logistic regression will outperform naive Bayes and achieve a lower asymptotic error rate.

[6] Dr. K. Usha Rani worked towards enhancing the classifier accuracy through various pre-processing techniques and ensemble techniques. A hybrid method was proposed to enhance the classification accuracy of Breast Cancer data sets. The data sets were preprocessed to remove missing values. The feature selection methods helped to eliminate the attributes that had no significance in the classification process. Bagging the training dataset

improved the decision tree. This combination gave higher classification accuracy of the selected data sets.

[7] Sacha explored miRNAs. microRNAs represent critical regulators of tumor cell differentiation, proliferation, cell cycle progression, invasion and metastasis. Based on microRNA arrays various miRNAs have been described as oncogenes or tumor suppressors and many of them are used for diagnosis and predictive tools. A new domain of gene selection was explored through the work.

[8] Pengyi Yang presented a review on Ensemble learning. These methods effectively deal with small sample size, high-dimensionality, and complexity data structures. The aim of this article is two-fold.

Dietterich TG worked to alleviate the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for overfitting the training data [9]. Random forests were used for high-dimensional datasets to generate multiple prediction models each with a different feature subset.

[10] Ben-Dor A worked on applying bagging and boosting methods to classify tumors using gene expression profiles. A comparative study of bagging and boosting with other individual classifiers such as *k*-nearest neighbors (*k*NN), clustering based classifiers, support vector machines (SVM), linear discriminant analysis (LDA), and classification trees was done. The conclusion was that ensemble methods of bagging and boosting performed similarly to other single classification algorithms included in the comparison. However, Dudoit *et al.* [11] later revealed that much better results can be achieved through minor tuning and modification.

Dettling and B¨uhlmann [12] proposed an algorithm called LogitBoost which replaced the exponential loss function used in AdaBoost with a log-likelihood loss function. The results

proved that LogitBoost is more accurate in classification of gene expression data compared to the original AdaBoost algorithm.

Long [13] proposed several customized boosting algorithms for microarray data classification. The experimental results indicated that the customized boosting algorithms performed favorably well compared to SVM-based algorithms. In comparison to the single tree classifier, Tan and Gilbert [14] demonstrated that, overall, ensemble methods of bagging and boosting are more robust and accurate in microarray data classification using seven publicly available datasets.

Geurts *et al.* [15] proposed a tree ensemble method called "extra-trees" which selected at each node the best among *k* randomly generated splits. This method was an improvement on random forests because unlike random forests which were grown with multiple subsets, the base trees of extra-trees were grown from the complete learning set and by explicitly randomizing the cut-points.

In the study of ensemble learning lastly, a trend to analyze microarray data at the pathway level was proposed [16]. The genes selected by random forests for sample classification were treated as informative genes while the error rate of random forests was used to evaluate the association between pathway and the disease of interests.

In gene function prediction, Guan *et al.* [17] introduced a meta-ensemble based on SVM. This meta-ensemble contained three "base classifiers". They were the ensemble of SVMs trained using bagging for each gene ontology (GO) term, the hierarchical bayesian combination of SVM classifiers, and the naïve bayes combination of SVM classifiers. The prediction of this meta-ensemble was made by selecting the best performing one on each GO term.

Anirban Mukhopadhyay [18] revealed that abnormal expression of some specific miRNAs often resulted in the development of cancer. Microarray datasets containing the

expression profiles of several miRNAs were used for identification of miRNAs which were differentially expressed in normal and malignant tissue samples. A multiobjective feature selection approach was proposed which used Genetic Algorithm for multiobjective optimization and support vector machine (SVM) classifier as a wrapper for evaluating the chromosomes that encode feature subsets.

Rasool Fakoor [19] proposed method over previous cancer detection approaches is the possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one. He brought forward two issues related to precise prediction of cancer. One was the high dimensionality of gene expression data and the other was the small sample size. He used a sparse autoencoder method to learn a concise feature representation from unlabeled data. His results provided a more general and scalable approach to deal with gene expression data across different cancer types.

3. DEFINITION OF THE PROBLEM

[20] The decision trees used for classification were easy to understand and analyze by the domain experts. However, the tree needed to be traversed repeatedly leading to higher complexity. Decision tree has been successfully applied for feature selection to analyze the performance of classifier. Using CART has enhanced the classification accuracy of a particular dataset.

SVM tries to find a hyper plane that separates the examples of different outcomes. Being primarily designed for two-class problems, SVMs find a hyper plane with a maximum distance to the closest point of the two classes; such a hyper plane is called the optimal hyper plane. A set of instances that is closest to the optimal hyper plane is called a support vector. The performance is more robust and less dependent on the specific selection of parameters. However, SVM is considered as black box.

Bayesian Networks can be used both as predictive and descriptive models. In prediction they constitute an efficient tool for solving different inference tasks. As a descriptive tool they possess the ability to efficiently represent the dependence/independence relationships among the random variables. However, the classes need to be mutually exclusive. Redundant attributes lead to misclassification.

None of these supervised learning algorithms are full proof in itself. Hence, we propose to define a combinational unsupervised learning algorithm. The mapping will be done amongst normal cells and tumour cells. This will help us build a trained data set. This trained data can then be used for testing the patient's gene dataset. The algorithm will work on miRNA data. Thus, our algorithm will help us select a specific gene expression from miRNA dataset to predict lung cancer.

4. OBJECTIVES AND ISSUES

The tumor develops slowly over a period of years. As cancer grows, it can cause new blood vessels to form. These vessels feed the cancer cells, allowing a tumor to grow. The cancer can then spread to other parts of the body.

There is no single test that can accurately diagnose cancer. The complete evaluation of a patient usually requires a thorough history and physical examination along with diagnostic testing. Diagnostic procedures for cancer may include imaging, laboratory tests (including tests for tumor markers), tumor biopsy, endoscopic examination, surgery, or genetic testing. The diagnosis also requires invasive biopsies in case of bone marrow cancer and esophagus cancer. The accuracy of clinical and pathological tests has a huge scope of improvement. Secondly, the delay is not acceptable as the tumour replicates at a rapid rate.

Cancer is a silent killer which creeps in the body without any symptoms. The clinical diagnostics are extremely taxing for the body and mind as well. Thus, we intend to reduce the burden on the patient in the following ways:

1) To enhance the prediction accuracy by using gene expression data.
2) To reduce high dimensionality of the dataset.
3) To find responsible miRNA biomarkers for lung cancer.

5. SCOPE OF STUDY

- Major prediction parameters dealing with the disease are [3]:

  (1) Occurrence of cancer,

  (2) Metastases and

  (3) Overall survival of the patient.

However, we restrict ourselves to predicting the presence of the disease.

- Various types of cancer have been studied.

Table 1: [26] Study of different types of cancer with respect to sex and age

American Cancer Society
Estimated New Cancer Cases by Sex and Age (Years), 2014

	All ages	Younger than 45	45 and Older	Younger than 65	65 and Older
All sites, men	855,220	53,990	801,230	370,880	484,340
All sites, women	810,320	87,920	722,400	390,910	419,410
Colon & rectum, men	71,830	3,680	68,150	30,160	41,670
Colon & rectum, women	65,000	3,260	61,740	22,820	42,180
Lung & bronchus, men	116,000	1,690	114,310	38,190	77,810
Lung & bronchus, women	108,210	2,020	106,190	34,410	73,800
Breast, women	232,670	25,500	207,170	133,310	99,360
Prostate	233,000	1,430	231,570	98,010	134,990

Projected cases are based on 1995-2010 incidence rates from 49 states and DC as reported by the North American Association of Central Cancer Registries (NAACCR).
Note: Estimates should not be compared with those from previous years because of ongoing changes in the method for estimating new cancer cases.

American Cancer Society, Surveillance Research, 2014

Table 2: [27] Study of different types of cancer with Indian Statistics

	India		United States	
	Male	Female	Male	Female
Cancer Rates, all sites except skin	99.0	104.4	361.4	283.2
Oral	12.8	7.5	6.3	3.7
Oesophagus	7.6	5.1	4.9	1.4
Stomach	5.7	2.8	7.3	3.6
Lung	9.0	2.0	58.6	34.0
Colon/Rectum	4.7	3.2	40.6	30.7
Breast	—	19.1	—	91.4
Ovary	—	4.9	—	10.6
Cervix	—	30.7	—	7.8
Endometrial	—	1.7	—	15.5
Prostate	4.6	—	104.3	—
Liver	2.3	2.0	4.2	1.7
Bladder	3.2	0.7	23.4	5.4
Kidney	1.2	0.5	11.2	6.0
Melanoma of the skin	0.3	0.2	4.2	1.7

However, we confine ourselves to Lung Cancer.

- miRNA Gene Expression

Genetic factors play an important role in prediction of lung cancer. Mutations in several genes have been linked to lung cancer. This suggests that there is a genetic component to lung cancer. Our study will be based on miRNA gene expression dataset.

6. METHODOLOGY, TOOLS AND TECHNIQUES

The proposed algorithm is explained as follows:

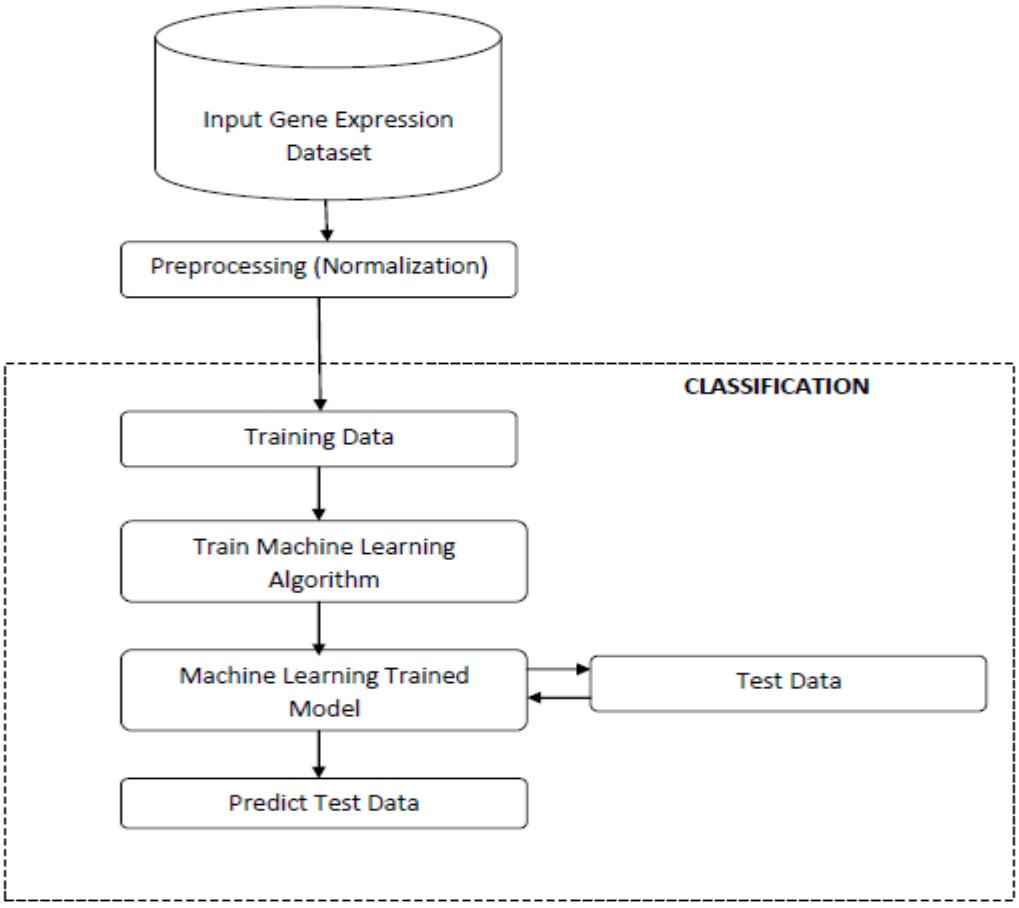

Figure 1: System Architecture

A gene expression dataset is provided as an input to the system. Preprocessing of data is done. Normalization concepts are used. This helps in reducing the dimensionality of the input dataset. Proper care will be taken to ensure that important data is not lost in the process. The trained data is then used as learning algorithm. This trained model is then used to test a new patient's gene expression. This will help to predict the occurrence of disease.

Gene Expression

Cell is the biological unit of living organisms. Human being are multi-cellular organisms called as Eukaryotic consisting of 100 trillion (10^14) cells. Cell constitutes of Deoxyribonucleic acid (DNA). These are arranged in two long strands that form a spiral called a double helix. The DNA transcription process leads to the formation of mRNA (messenger Ribonucleic acid) which conveys genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression.

[29] The first human disease known to be associated with miRNA deregulation was chronic lymphocytic leukemia. miRNAs modulate gene expression impact all cell functions, cell cycle and metabolism. Specific miRNA signatures have been associated with distinct subsets of solid tumors and hematological malignancies [21]. Various miRNAs have been described as oncogenes or tumor suppressors and many of them are used for diagnosis and as prognostic or predictive tools [22]. microRNAs (miRNAs) is particularly fascinating because of the breadth of their interactions facilitated by their synergistic/combinatorial relationships with target genes. [23] miRNAs are pivotal regulators of development and cellular homeostasis through their control of diverse biological processes.

Deep Learning [24]

Deep learning algorithms are based on distributed representations. Deep learning helps to disentangle the abstractions at higher level and pick out which features are useful for learning. Many deep learning algorithms are framed as unsupervised learning problems. These algorithms make use of the unlabeled data, which is usually more abundant than labeled data, making this an important benefit of these algorithms. These algorithms can model complex non-linear phenomenon as genomic types. It learns a distributed and hierarchical feature representation and can exploit unlabeled data well.

In a deep learning scheme each layer is treated separately and successively trained in a greedy manner: once the previous layers have been trained, a new layer is trained from the encoding of the input data by the previous layers. [25] Deep learning methods achieve very good accuracy for tasks where a large set of data is available, even if only a small number of instances are labeled.

Some models include deep neural networks, which is a shallow Artificial Neural Network and can deal with complex linear relations. A deep belief network (DBN) is a probabilistic, generative model made up of multiple layers of hidden units. A DBN can be efficiently trained in an unsupervised, layer-by-layer manner where the layers are typically made of restricted Boltzmann machines (RBM).

Tools and Techniques

R is a free software programming language and software environment for statistical computing and graphics [30]. R provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages.

Input: Training data

PID	GEO asscession number	lymph node status	time to relapse or last follow-up (months)	relapse (1=True)	ER Status	Brain relapses (1=yes, 0=no)
3	GSM36793	negative	101	0	ER-	0
5	GSM36796	negative	118	0	ER+	0
6	GSM36797	negative	9	1	ER-	0
7	GSM36798	negative	106	0	ER-	0

8	GSM36800	negative	37	1	ER-	0
….						
….						
….						
900	GSM36887	negative	108	0	ER+	0
901	GSM36889	negative	108	0	ER+	0
903	GSM36892	negative	110	0	ER+	0
909	GSM36894	negative	109	0	ER+	0
913	GSM36911	negative	80	1	ER+	0

Output: Test data for Breast cancer

Confusion table

	NoRelapse	Relapse
NoRelapse	102	18
Relapse	50	17

Sensitivity = 25.3731343283582

Specificity = 85

overall accuracy = 63.6363636363636

overall error rate = 36.3636363636364

18 NoRelapse misclassified as Relapse

50 Relapse misclassified as NoRelapse

AUC = 0.62363184079602

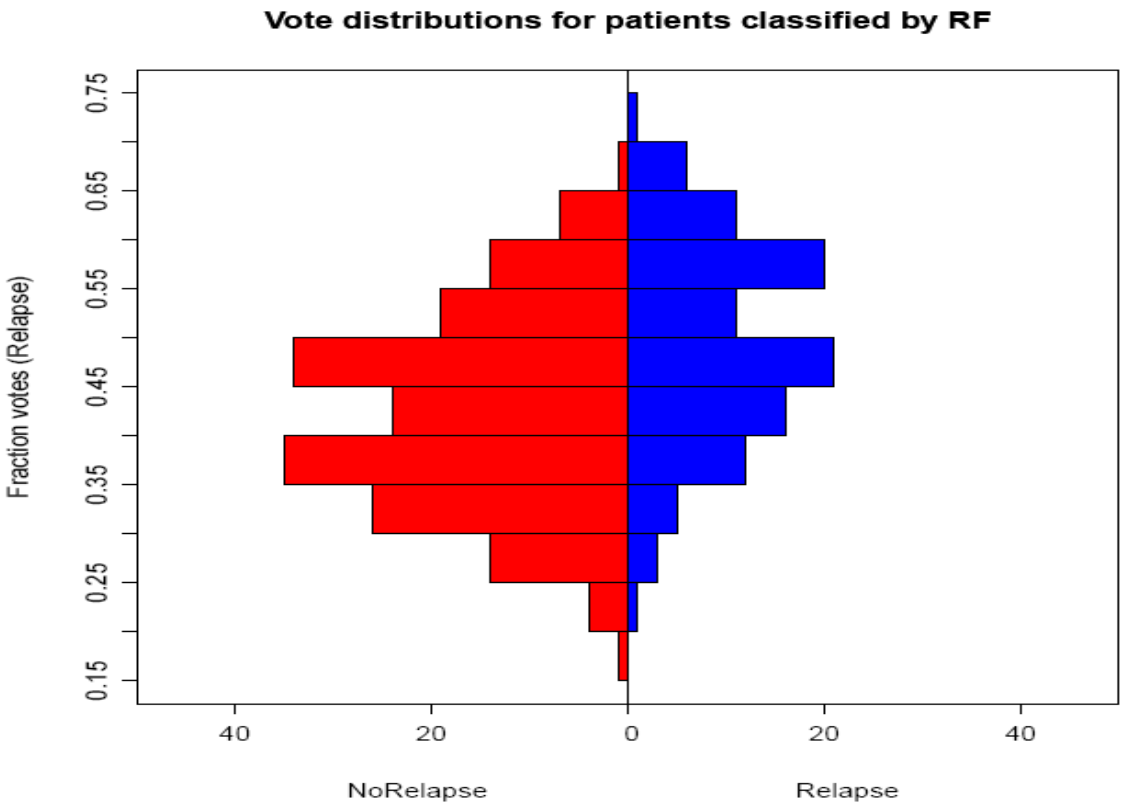

Figure 2: Graph to show the ratio of relapse and no relapse in breast cancer

7. PLAN OF RESEARCH WORK

The proposed research work is planned to be carried out in thirty six months. For close monitoring, we propose to divide this period in six semesters of six months each. The broader outline of the work to be carried out in these semesters is as under:

- **Review of literature**: Literature review is an important aspect of an organized research work. We plan to continue the literature review of existing works throughout the Semesters I to V.
- **Development of deep learning algorithm using miRNA for gene selection to predict lung cancer**: We plan to develop algorithm for predicting lung cancer through gene expression. We plan to continue this aspect from Semester I to V.
- **Comparison of our algorithm with numerical results**: The designed method needs to be analyzed with previous methods to check the accuracy and efficiency. We plan to compare our method throughout Semester II to V.
- **Communication of research work in Journals and Conferences:** We Plan to communicate all the research work carried out in this study to international Journals and Conferences during Semester I to V.
- **Writing of Thesis**: The research Work is planned to be organized in thesis consisting of five chapters. This will be done in the V th and VI th Semesters.

Time line Chart indicating the detailed work plan:

Time Activity	Sem I	Sem II	Sem III	Sem IV	Sem V	Sem VI
Literature Review						
Development of deep learning algorithm using miRNA gene expression						
Comparison of deep learning algorithm using miRNA gene expression with numerical results						
Communication of research work in Journals and Conferences						
Thesis writing						

8. REFERENCES

1. Anirban Mukhopadhyay, "An SVM-Wrapped Multiobjective Evolu-tionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers" IEEE Transactions On Nanobioscience, Vol. 12, No. 4, December 2013.
2. Joseph A. Cruz and David S. Wishart "Applications of Machine Learning in Cancer Prediction and Prognosis" Cancer Inform. 2006; 2: 5977. Published online Feb 11, 2007. PMCID: PMC2675494
3. Stiphout ab E.O. Postma c V. Valentini d P. Lambina "The contribution of machine learning to predicting cancer outcome" R.G.P.M.van
4. Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naïve Bayes and J48 Classi_cation Algorithm for Data Classi_cation", International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.
5. John Halloran, "Classi_cation: Naive Bayes vs Logistic Regression", December 21, 2009
6. D.Lavanya and Dr.K.Usha Rani, "Ensemble Decision Tree Classier For Breast Cancer Data", International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012.
7. Sacha I Rothschild, "microRNA therapies in cancer", Molecular and Cellular Therapies 2014.
8. Pengyi Yang, Yee Hwa Yang, Bing B. Zhou1 and Albert Y. Zomaya, A review of ensemble methods in bioinformatics, Bioinformatics, 5, 2010.
9. Dietterich TG. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization". Machine Learning. 2000; 40:139158.

10. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. "Tissue Classi_cation with Gene Expression Pro_les", Journal of Computational Biology. 2000
11. Dudoit S, Fridlyand J, Speed T., "Comparison of discrimination methods for the classi_cation of tumors using gene expression data", Journal of the American Statistical Association. 2002; 97:77 87.
12. Dettling M, Buhlmann P.," Boosting for tumor classi_cation with gene expression data", Bioinformatics. 2003; 19(9):10611069.
13. Long P., "Boosting and Microarray Data", Machine Learning. 2003; 53:3144.
14. Tan A, Gilbert D., "Ensemble machine learning on gene expression data for cancer classi_cation", Applied Bioinformatics. 2003; 2(3 Suppl):S75S83.
15. Geurts P, Fillet M, Seny D, Meuwis M, Malaise M, Merville M, etal., "Proteomic mass spectra classi_cation using decision tree based ensemble methods", Bioinformatics. 2005; 21(15):31383145.
16. Curtis RK, Oresic M, Vidal-Puig A., "Pathways to the analysis of microarray data", TRENDS in Biotechnology. 2005; 23(8):429435.
17. Guan Y, Myers C, Hess D, Barutcuoglu Z, Caudy A, Troyanskaya O., " Predicting gene function in a hierarchical context with an ensemble of classi_ers", Genome Biology. 2008; 9:S3.
18. Anirban Mukhopadhyay and Ujjwal Maulik, "An SVM-Wrapped Multi-objective Evolutionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers", IEEE Transactions On Nanobioscience, Vol. 12, NO. 4, December 2013 275.
19. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber, "Using deep learning to enhance cancer diagnosis and classi_cation", Bioin-formatics 2014.

20. S.Kharya, D. Dubey and S. Soni, " Predictive Machine Learning Techniques for Breast Cancer Detection", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (6), 2013, 1023-1028
21. He L, Hannon G. "MicroRNAs: small RNAs with a big role in gene regulation". Nature 5 (7), July 2004, 522531.
22. Artmann S, Jung K, Bleckmann A, Beissbarth T., "Detection of simultaneous group e_ects in microRNA expression and related target gene sets". In Provero, Paolo. PLoS ONE 7 (6): e38365, 2012.
23. Bing Liu, Jiuyong Li and Murray J.Cairns, "Identifying miRNAs, targets and functions", Briefings in Bioinformatics. Vol 15. November 2012.
24. Yoshua Bengio and Arnaud Bergeron, "Deep Learners Benet More from Out-of-Distribution Examples", Proceedings of the 14thInternational Con-ference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15.
25. Ludovic Arnold, Sbastien Rebecchi, Sylvain Chevallier, Hlne Paugam- Moisy, "An Introduction to Deep Learning", Esann 2011.
26. American Cancer Society, Surveillance Research, 2014
27. Cancer and Diet in India, Sinha R,Anderson DE. www.jpgmonline.com
28. www.seer.cancer.gov
29. Wikipedia : miRNA
30. Wikipedia : R analyzing tool